4
Parameter Estimation

The estimation of the parameter vector $\boldsymbol{\theta}$ is carried out by maximizing the quasi-likelihood of the density function in (3-1). In a more general framework we cannot suppose that our probability model is correctly specified, so we use the Quasi-Maximum Likelihood Estimator (QMLE), which is the same as the Maximum Likelihood Estimator under the correct specification. Thus, we can write the conditional quasi-likelihood based on a sample $\{y_t\}_{t=1}^{T}$ as

$$\mathcal{L}_T(\boldsymbol{\theta}) = \sum_{t=1}^{T} \log \left[\sum_{i \in \mathbb{T}} B_i(\mathbf{x}_t; \boldsymbol{\theta}_i) \pi(y_t | \mathbf{x}_t; \boldsymbol{\psi}_i) \right]. \tag{4-1}$$

Numerical optimization is carried out using the EM algorithm of (19), as shown in appendix A. The idea behind the EM algorithm is to maximize a sequence of simple functions which leads to the same solution as maximizing a complex function. This technique were also used by (36), (38), Wong and Li (1999,2000), (31) and (9), among others.

4.1
Asymptotic Theory

In this section we present a set of asymptotic results with respect to the estimator. First, we present a set of assumptions about the (unknown) true probability model.

Assumption 1 *The observed data are a realization of a stochastic process* $\{(y_t, \mathbf{x}_t)\}_{t=1}^{T}$*, where the unknown true probability model* $\mathcal{G}_t \equiv \mathcal{G}[(y_t, \mathbf{x}_t); \cdot]$ *is a continuous density on* $\mathbb{R}$*, and the true likelihood function is identifiable and has a unique maximum at* $\boldsymbol{\theta}_0$*.*

We define $\boldsymbol{\theta}^*$ as the parameter vector that minimize the Kullback-Leibler divergence criterion between the true probability model, $\mathcal{G}_t$, and the estimated probability model, $f(\cdot; \boldsymbol{\theta})$. Hence, the QMLE $\widehat{\boldsymbol{\theta}}_T$ of $\boldsymbol{\theta}^*$, is defined as:

$$\widehat{\boldsymbol{\theta}}_T = \underset{\boldsymbol{\theta} \in \boldsymbol{\Theta}}{\mathsf{argmax}} \mathcal{L}_T(\boldsymbol{\theta}), \tag{4-2}$$

where $\mathcal{L}_T(\boldsymbol{\theta})$ is established in (4-1). Let $\#$ be the cardinality operator.

Assumption 2 *The parameter vector $\boldsymbol{\theta}^*$ is interior to a compact parameter space $\boldsymbol{\Theta} \in \mathbb{R}^{r_1} \times \mathbb{R}_+^{r_2}$, where $r_1 = 2(\#\mathbb{J}) + (p+1)(\#\mathbb{T})$ and $r_2 = \#\mathbb{T}$.*

The identifiability of mixture of experts models was shown in (35) for the case where the gating functions are multinomial logits. Since our gating function is different, the conditions presented there are not adequate. We show in Appendix B that under mild conditions, the model is identifiable such that the following assumption holds.

Assumption 3 *The tree mixture-of-expert structure, as presented in (3-1), is identifiable, in the sense that, for a sample $\{y_t; \mathbf{x}_t\}_{t=1}^T$, and for $\boldsymbol{\theta}_1$, $\boldsymbol{\theta}_2 \in \boldsymbol{\Theta}$,*

$$\prod_{t=1}^{T} f(y_t|\mathbf{x}_t;\boldsymbol{\theta}_1) = \prod_{t=1}^{T} f(y_t|\mathbf{x}_t;\boldsymbol{\theta}_2) \quad , a.s.$$

is equivalent to $\boldsymbol{\theta}_1 = \boldsymbol{\theta}_2$.

The following theorem establishes the existence of the QMLE.

Teorema 4.1 (Existence) *Under Assumptions 1 – 3, the QMLE exists and $\mathbb{E}[\mathcal{L}_T(\boldsymbol{\theta})]$ has a unique maximum at $\boldsymbol{\theta}^*$.*

To ensure the consistency of the QMLE, we state additional conditions.

Assumption 4 *The process $\{(y_t, \mathbf{x}_t)\}_{t=1}^T$ is strictly stationary and strong mixing.*

Assumption 5 *Let $\mathbf{Y}_t = (y_t,\, \mathbf{x}_t')'$, then $\mathbb{E}\left[\mathbf{Y}_t\mathbf{Y}_t'\right] < \infty$.*

Teorema 4.2 *Under Assumptions 1–6, $\widehat{\boldsymbol{\theta}}_T \overset{a.s.}{\to} \boldsymbol{\theta}^*$.*

For asymptotic normality we need the following additional assumption:

Assumption 6 $\mathbb{E}\left[\mathbf{Y}_t \otimes \mathbf{Y}_t \otimes \mathbf{Y}_t \otimes \mathbf{Y}_t\right] < \infty$

Teorema 4.3 (Asymptotic Normality) *Under Assumptions 1–6,*

$$\sqrt{T}\left(\widehat{\boldsymbol{\theta}}_T - \boldsymbol{\theta}^*\right) \overset{D}{\to} \mathsf{N}\left(0, \mathbf{A}(\boldsymbol{\theta}^*)^{-1}\mathbf{B}(\boldsymbol{\theta}^*)\mathbf{A}(\boldsymbol{\theta}^*)^{-1}\right),$$

where

$$\mathbf{A}(\boldsymbol{\theta}^*) = \mathbb{E}\left[-\frac{\partial^2\mathcal{L}_T(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}\partial\boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}^*}\right] \text{ and } \mathbf{B}(\boldsymbol{\theta}^*) = \mathbb{E}\left[\frac{\partial\mathcal{L}_T(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}}\bigg|_{\boldsymbol{\theta}^*}\frac{\partial\mathcal{L}_T(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}^*}\right].$$